TEMA 0: conceptos y resultados básicos

José R. Berrendero

Departamento de Matemáticas, Universidad Autónoma de Madrid

Temas a tratar

- Variables y vectores aleatorios.
- Métodos de estimación:
  - Momentos
  - Máxima verosimilitud
  - Estimadores bayesianos
- Criterios para valorar un estimador
  - Distribución en el muestreo
  - Sesgo y varianza
  - Criterios asintóticos. Ley de los grandes números y teorema central del límite
- Otras técnicas de inferencia: intervalos de confianza y contrastes de hipótesis.

Variables aleatorias

Una **variable aleatoria** (v.a.) X es una función (medible) $X : \Omega \to \mathbb{R}$ que asigna un número real $X(\omega)$ a cada resultado ω de un experimento aleatorio.

La **función de distribución** $F : \mathbb{R} \to [0, 1]$ de una v.a. X es

$$F(x) = \mathrm{P}(X \leq x)$$

- Si X es el número de caras al tirar una moneda dos veces, determina su función de distribución.
- La funcion de distribución determina completamente la distribución de una v.a.
- F es monótona no decreciente, continua por la derecha, $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$

Variables discretas y continuas

Una v.a. es **discreta** si toma un número numerable de valores $x_1, x_2, \ldots$. Su función de probabilidad es

$$p(x_i) = p_i = P(X = x_i), \quad i = 1, 2, \ldots$$

Una v.a. es **continua** si existe una función $f(x) \geq 0$ con $\int f(x)dx = 1$ tal que

$$\mathrm{P}(a < X < b) = \int_a^b f(x)dx.$$

- Se dice que f es la **función de densidad** de X.
- Se verifica $F(x) = \int_{-\infty}^{x} f(t)dt$, y $F'(x) = f(x)$ si F es derivable es x.

Distribuciones de probabilidad con R

- Cada distribución se nombra mediante una palabra clave o *alias*.
- Una lista completa de las distribuciones: `help(Distributions)`
- A cada distribución se le antepone un prefijo que determina una función relacionada con ella:

Funciones	Prefijos
Función de distribución	`p`
Función cuantílica	`q`
Función de densidad (continuas) o de probabilidad (discretas)	`d`
Generación de números aleatorios	`r`

Distribuciones de probabilidad con R

Algunas cuestiones:

- Calcula la probabilidad de que una variable normal estándar sea mayor que 1
- Determina un valor x tal que $P(Z > x) = 0.3$, donde Z es una v.a. con distribución normal estándar (es decir, el percentil 70 de la distribución)
- Calcula la mediana de una variable exponencial de media 2
- Genera 100 valores de una distribución normal y representa el correspondiente histograma usando ggplot2
- Añade al gráfico la curva de densidad teórica de los datos generados

Distribuciones de probabilidad con R

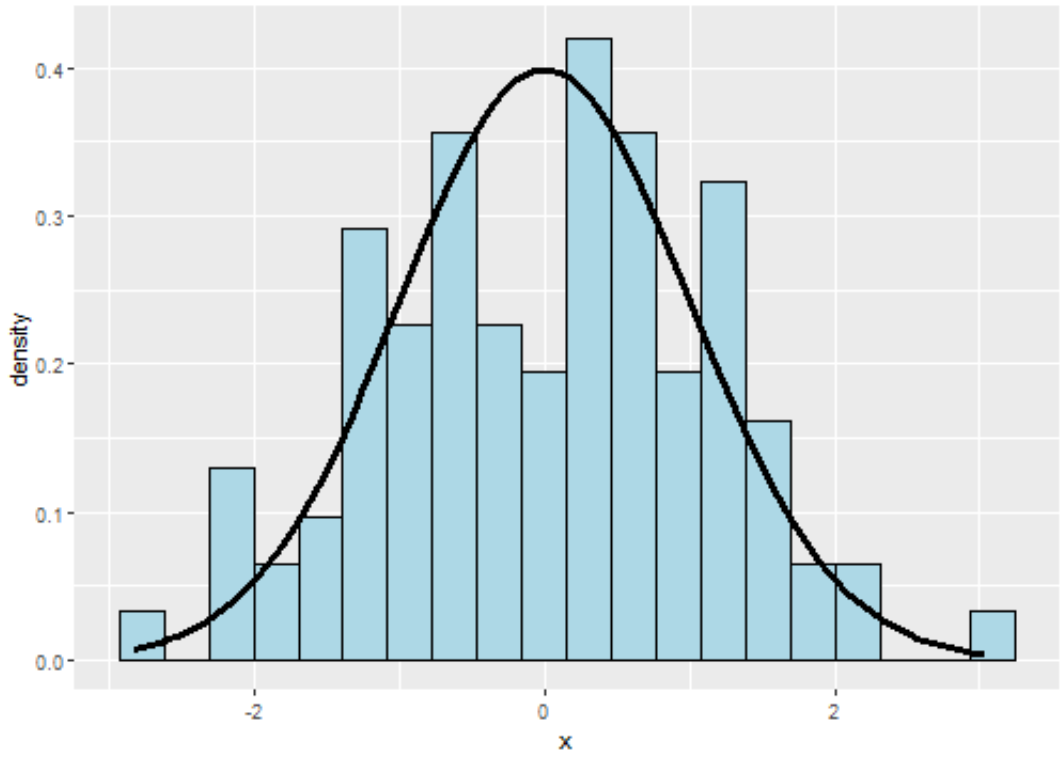
```
library(tidyverse)
# P(Z > 1)
1 - pnorm(1)

# P(X > x) = 0.3
qnorm(0.7)

# Mediana de exponencial de media 2
qexp(0.5, rate=0.5)

# Números aleatorios
x <- rnorm(100)
df <- data.frame(x=x)
ggplot(df) +
  geom_histogram(aes(x=x, y=..density..),
                 col='black', fill='lightblue', bins=20)
  geom_function(fun = dnorm, col='black', size=1.3)
```

Distribuciones de probabilidad con R

```
## [1] 0.1586553
```

```
## [1] 0.5244005
```

```
## [1] 1.386294
```



Esperanza

La **esperanza** (o media) de una v.a. X es $\mathrm{E}(X) := \mu = \int x dF(x)$. Esta es una notación que significa:

- $\mathrm{E}(X) = \sum_i x_i \mathrm{P}(X = x_i)$, si X es discreta.
- $\mathrm{E}(X) = \int x f(x) dx$, si X es continua.
- Es un resumen del centro en torno al cual toma valores X
- Se puede dar un significado más preciso a la notación $\int x dF(x)$
- Si $Y = g(X)$, entonces $\mathrm{E}(Y) = \int g(x) dF(x)$. Esto significa que no es necesario calcular la distribución de Y, basta conocer la de X.
- La esperanza es lineal $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$

Varianza y covarianza

Sea X una v.a. con esperanza μ.

La **varianza** de X es $\mathrm{Var}(X) := \sigma^2 = \mathrm{E}[(X-\mu)^2]$.

La **desviación típica** de X es σ.

- La varianza mide la dispersión de los valores que toma X
- **No** es lineal: $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$

Sean X e Y v.a. con esperanzas μ_x y μ_y, y varianzas σ_x^2 y σ_y^2:

Se define la **covarianza** entre X e Y como
$\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mu_x)(Y-\mu_y)]$.

El **coeficiente de correlación** entre X e Y es

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}$$

- La correlacion mide el grado de relación lineal entre X e Y
- Siempre $-1 \leq \rho(X,Y) \leq 1$.

Algunas desigualdades

Desigualdad de Markov: Sea X una v.a. no negativa con esperanza $\mathrm{E}(X) < \infty$. Entonces, para todo $\epsilon > 0$,

$$\mathrm{P}(X \geq \epsilon) \leq \frac{\mathrm{E}(X)}{\epsilon}.$$

Desigualdad de Chebychev: Sea X una v.a. con esperanza μ y varianza σ^2. Entonces, para todo $\epsilon > 0$,

$$\mathrm{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Desigualdad de Cauchy-Schwarz: Dadas dos v.a. X e Y con varianzas finitas,

$$\mathrm{E}|XY| \leq \sqrt{\mathrm{E}(X^2)\mathrm{E}(Y^2)}.$$

Desigualdad de Jensen: Si g es una función convexa,

$$\mathrm{E}[g(X)] \geq g[\mathrm{E}(X)].$$

Vectores aleatorios

Se dice que $X = (X_1, \ldots, X_p)$ es un **vector aleatorio** si X_i es una v.a. para todo $i = 1, \ldots, p$. Las distribuciones de las coordenadas X_i se llaman **distribuciones marginales**.

- El vector aleatorio X es continuo si existe una función de densidad conjunta $f : \mathbb{R}^p \to [0, \infty]$ tal que

$$\mathrm{P}(X \in A) = \int_A f(x) dx,$$

  para $A \subset \mathbb{R}^p$ (medible).

- Las marginales son independientes si

$$\mathrm{P}(X_1 \in A_1, \ldots, X_p \in A_p) = \mathrm{P}(X_1 \in A_1) \cdots \mathrm{P}(X_p \in A_p)$$

- En el caso continuo la independencia equivale a que la densidad conjunta es el producto de las p densidades marginales.

- Bajo independencia $\mathrm{E}(X_1 \cdots X_p) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_p)$

- Si X e Y son independientes y tienen varianzas finitas, $\mathrm{Cov}(X, Y) = 0$. El recíproco **no** es cierto.

Vectores aleatorios

Si $X = (X_1, \ldots, X_p)'$ un vector aleatorio p-dimensional:

- Su **vector de medias** es $\mathrm{E}(X) = \mu = (\mu_1, \ldots, \mu_p)'$, donde $\mu_i = \mathrm{E}(Y_i)$
- Su **matriz de covarianzas** es $\mathrm{Var}(X) := \mathrm{Cov}(X) := \Sigma$, cuya posición (i, j) es $\sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$.
- Es fácil comprobar

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu)(X - \mu)'] = \mathrm{E}(XX') - \mu\mu'.$$

- Si A es una matriz $q \times p$ y $b \in \mathbb{R}^q$, entonces $\mathrm{E}(AX + b) = A\mu + b$ y $\mathrm{Var}(AX + b) = \mathrm{E}[A(X - \mu)(X - \mu)'A'] = A\Sigma A'$.
- Da una expresión general de $\mathrm{Var}(X_1 \mp X_2)$
- Si $X_1, \ldots, X_p$ son independientes, ¿cuanto vale $\mathrm{Var}(a_1 X_1 + \cdots + a_p X_p)$?

Esperanza y varianza condicionadas

Sean X e Y dos v.a. La esperanza y la varianza de Y condicionadas a X (es decir, dado que conocemos X) se definen:

$$\mathrm{E}(Y|X) = \int x dF_{Y|X}(x), \;\; \mathrm{Var}(Y|X) = \mathrm{E}[(Y - \mathrm{E}(Y|X))^2|X]$$

- Las distribuciones condicionadas se calculan de manera similar a las probabilidades condicionadas. Por ejemplo, si f es la densidad conjunta de (X, Y) y f_X la densidad marginal de X,

$$f_{Y|X}(y) = \frac{f(x,y)}{f_X(x)}.$$

- Tanto $\mathrm{E}(Y|X)$ como $\mathrm{Var}(Y|X)$ son variables aleatorias (funciones de X).

- Si X e Y son independientes $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

Esperanza y varianza condicionadas

- $\mathrm{E}(g(X)Y|X) = g(X)\,\mathrm{E}(Y|X)$
- $\mathrm{E}(a_1Y_1 + a_2Y_2|X) = a_1\mathrm{E}(Y_1|X) + a_2\mathrm{E}(Y_2|X)$
- Ley de la esperanza iterada: $\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|X)]$
- El valor $\mathrm{E}(Y|X)$ da la mejor predicción posible de Y a partir de X: para cualquier g,

$$\mathrm{E}[(Y - g(X))^2] \geq \mathrm{E}[(Y - \mathrm{E}(Y|X))^2]$$

- Una identidad útil para la varianza:

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)]$$

Distribución normal multivariante

El vector aleatorio X es **normal p-dimensional** con vector de medias μ y matriz de covarianzas Σ si su densidad es

$$f(x) = |\Sigma|^{-1/2}(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$

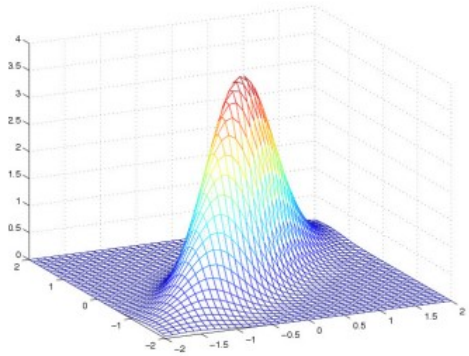
- ¿Qué resulta si $\mu = 0$ y $\Sigma = \mathbb{I}$?
- Usamos la notación $X \equiv \mathrm{N}_p(\mu, \Sigma)$.
- Si $X \equiv \mathrm{N}_p(\mu, \Sigma)$, A es matriz $q \times p$ y $b \in \mathbb{R}^q$, entonces

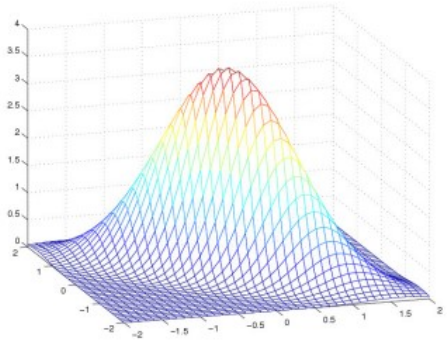
$$AX + b \equiv \mathrm{N}_q(A\mu + b, A\Sigma A').$$

- Si $X \equiv N_p(\mu, \Sigma)$. ¿Cuál es la distribución de $\Sigma^{-1/2}(X - \mu)$?

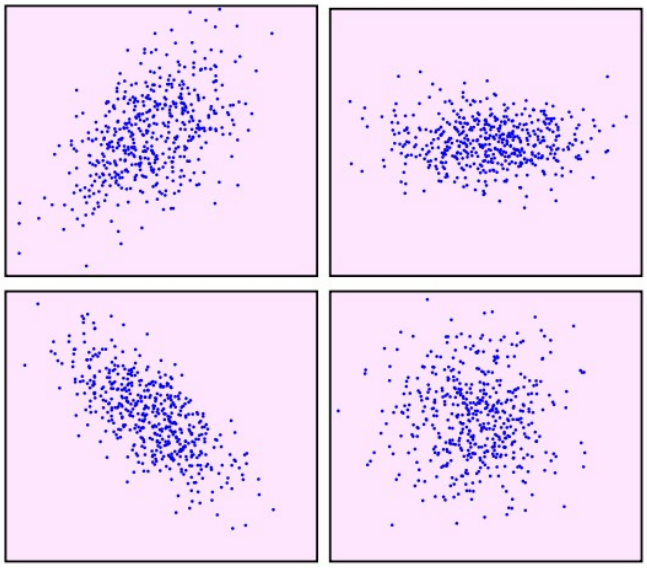
Distribución normal multivariante

$\mu = (0,0)'$ y $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ $\mu = (0,0)'$ y $\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$





Distribución normal multivariante



$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

Distribución normal multivariante

Sea $X \equiv N_p(\mu, \Sigma)$. Consideramos la partición $X = (X_a, X_b)$, con $X_a \in \mathbb{R}^q$ y $X_b \in \mathbb{R}^{p-q}$, y consideramos las particiones correspondientes de μ y Σ,

$$\mu = (\mu_a, \mu_b), \quad \Sigma = \left(\begin{array}{c|c} \Sigma_{aa} & \Sigma_{ab} \\ \hline \Sigma_{ba} & \Sigma_{bb} \end{array} \right),$$

- $X_a \equiv \mathrm{N}_q(\mu_a, \Sigma_{aa})$
- X_1 y X_2 son independientes si y solo si $\Sigma_{ab} = 0$.
- Distribución condicionada:

  $$X_a|(X_b = x) \equiv N_q(\mu_a - \Sigma_{ab}\Sigma_{bb}^{-1}(x - \mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}).$$

- Aplica la propiedad anterior a un vector normal bidimensional (X, Y) para obtener la distribución condicionada $Y|X$.

Modelos estadísticos

La muestra

Sea $X_1, \ldots, X_n$ un conjunto de n variables (o vectores, o funciones) aleatorias. En este curso suponemos $X_1, \ldots, X_n$ i.i.d. con distribución F

Modelo

Formular un modelo estadístico es especificar un conjunto de distribuciones al que pertenece F.

- Un modelo es **paramétrico** si cada distribución del conjunto es totalmente conocida salvo por el valor de un parámetro $\theta \in \mathbb{R}^d$.

$$F \in \{F_\theta : \theta \in \Theta \subset \mathbb{R}^d\}.$$

- **Identificabilidad**: si $\theta \neq \theta'$, entonces $F_\theta \neq F_{\theta'}$.
- **Modelos no paramétricos**: por ejemplo,

$$F \in \{F : F \text{ tiene función de densidad } f\}.$$

Método de momentos

Sean $X_1, \ldots, X_n$ vaiid con distribución determinada por $f(\cdot; \theta)$, donde $\theta = (\theta_1, \ldots, \theta_d)$ es un parámetro d-dimensional.

Los momentos $\alpha_k(\theta) := \mathrm{E}_\theta(X^k)$, $k = 1, \ldots, d$, son funciones de los los parámetros θ_i.

Para estimar θ el método de momentos consiste en resolver en $\theta_1, \ldots, \theta_d$ el sistema de ecuaciones

$$m_1 = \alpha_1(\theta) \ , \ \ldots \ , \ m_d = \alpha_d(\theta),$$

donde $m_k = \dfrac{\sum_{i=1}^n X_i^k}{n}$.

La idea es estimar el parámetro como aquel valor de θ que hace que los momentos poblacionales coincidan con los correspondientes momentos muestrales.

Ejemplos

- Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución uniforme en $(0, \theta)$. ¿Cuál es el estimador de momentos de θ?
- Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución con densidad $f(x; \theta) = (\theta + 1)x^{\theta}$, con $x \in [0, 1]$, $\theta > -1$. ¿Cuál es el estimador de momentos de θ?
- Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución $\mathrm{N}(\mu, \sigma^2)$. ¿Cuáles son los estimadores de momentos de μ y σ^2?

Método de máxima verosimilitud

En una urna cerrada hay 4 bolas, θ de ellas son blancas y $4-\theta$ son negras (con θ desconocido). Se llevan a cabo dos extracciones de bolas con reemplazamiento. Si tuviéramos que apostar por cuántas bolas blancas hay, ¿por qué valor apostaríamos dados los resultados obtenidos?

θ	$\mathrm{P}\{x_1 = B, x_2 = N\}$
0	0
1	3/16
2	4/16
3	3/16
4	0

Método de máxima verosimilitud

- Sea $x_1, \ldots, x_n$ una realización de una muestra $X_1, \ldots, X_n$ con función de densidad o de probabilidad conjunta $g(x_1, \ldots, x_n; \theta)$ en $(x_1, \ldots, x_n)$, donde $\theta \in \Theta \subset \mathbb{R}^d$.
- La **función de verosimilitud** $L : \Theta \to \mathbb{R}$ se define como

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = g(x_1, \ldots, x_n; \theta).$$

- Si $X_1, \ldots, X_n$ son vaiid con densidad o probabilidad $f(x; \theta)$, entonces $L(\theta) = \prod_{i=1}^n f(x_i; \theta)$
- Un **estimador de máxima verosimilitud (EMV)** de θ es un valor $\hat{\theta} \in \Theta$ tal que

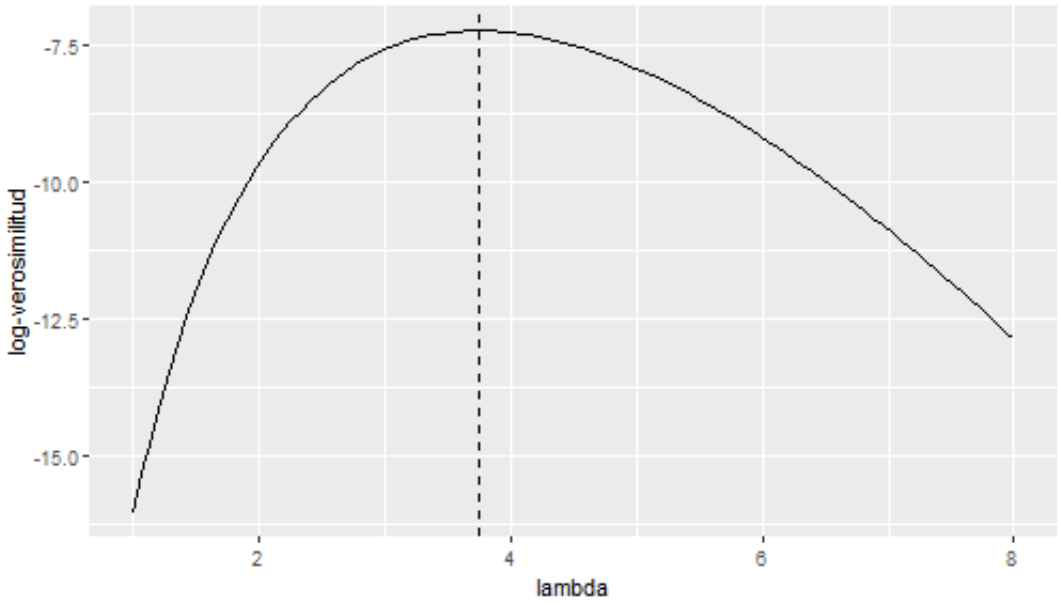
$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

En vez de maximizar directamente $L(\theta)$ suele resultar más conveniente maximizar
$\ell(\theta; x) = \ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(x_i, \theta)$.

Ejemplos

El número de llamadas por hora que se reciben en una centralita telefónica por hora sigue una distribución de Poisson de parámetro λ. En cuatro horas diferentes se han observado de forma independiente los siguientes números de llamadas: $x_1 = 2$, $x_2 = 3$, $x_3 = 5$, $x_4 = 5$. Calcula el EMV de λ.

$$\ell(\lambda) = -4\lambda + 15 \log \lambda - c, \quad \hat{\lambda} = 15/4.$$



- En general, $\hat{\lambda} = \bar{x}$ es el EMV de λ en una distribución de Poisson.

Ejemplos

- Sea $X_1, \ldots, X_n$ una muestra de vaiid de una distribución exponencial de parámetro θ. Calcula el EMV de θ.
- El tiempo de vida de los ratones con cierta enfermedad sometidos a un tratamiento es una v.a. con distribución exponencial de parámetro θ. Se lleva a cabo un experimento con n ratones, se observa el tiempo de vida de m de ellos, $x_1, \ldots, x_m$, pero se interrumpe el experimento transcurrido un tiempo T de manera que de los $n - m$ restantes solo se sabe que su tiempo de vida es superior a T. Calcula el EMV de θ. Se supone que todos los tiempos son independientes.
- Supongamos que $X_1, \ldots, X_n$ es una muestra de vaiid de una distribución uniforme en $(0, \theta)$. ¿Cuál es el EMV de θ?
- Sea $X_1, \ldots, X_n$ una muestra de vaiid de una distribución normal de media μ y varianza σ^2. Calcula el EMV de μ y de σ^2.

Métodos bayesianos

Dos interpretaciones de la probabilidad de un suceso:

- **Frecuentista**: el límite de la frecuencia relativa de veces que ocurre este suceso cuando un experimento aleatorio se va repitiendo más y más veces.
- **Bayesiana**: grado de creencia subjetiva en que tal suceso ocurra.

Si adoptamos un enfoque bayesiano, tiene sentido describir la incertidumbre sobre el parámetro mediante una distribución de probabilidad definida en el espacio paramétrico Θ.

Métodos bayesianos

El método bayesiano opera de la forma siguiente:

- Se establece una distribución a priori sobre Θ, $\pi(\theta)$, previa a la observación de la muestra que refleja la opinión de un experto sobre los valores del parámetro.
- La información sobre θ contenida en $\pi(\theta)$ se combina con el modelo estadístico para los datos $f(x|\theta)$ mediante el teorema de Bayes para calcular la llamada **distribución a posteriori** $\pi(\theta|x)$:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

- Se toma como estimador de θ alguna medida numérica de posición que resuma la distribución a posteriori. Usualmente la media o la moda.

Ejemplos

El número de llegadas por hora a las emergencias de un hospital sigue una distribución de Poisson de parámetro λ. A priori se supone que el valor de λ toma alguno de los siguientes valores con las siguientes probabilidades:

Valores	3	3.5	4	4.5	5
Probabilidades	0.1	0.2	0.4	0.2	0.1

Se han registrado las visitas durante $n = 10$ periodos de una hora independientes y el número total de visitas ha sido de 31. ¿Cuál es el estimador bayesiano de λ?

Ejemplos

- Sea $X_1, \ldots, X_n$ una muestra de vaiid de una distribución $\mathrm{B}(1, \theta)$. Queremos estimar la probabilidad de éxito θ. Se supone que la distribución a priori de θ es una distribución beta de parámetros α y β adecuados. Calcula $\hat{\theta} = \mathrm{E}(\theta|x)$.

- Sea $X_1, \ldots, X_n$ una muestra de vaiid de una distribución $\mathrm{N}(\mu, \sigma^2)$. Se supone que la distribución a priori de μ es también una distribución normal, $\mathrm{N}(\mu_0, \sigma_0^2)$. Comprueba que la distribución a posteriori es de nuevo normal, $\mathrm{N}(\mu_1, \sigma_1^2)$, donde

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

y

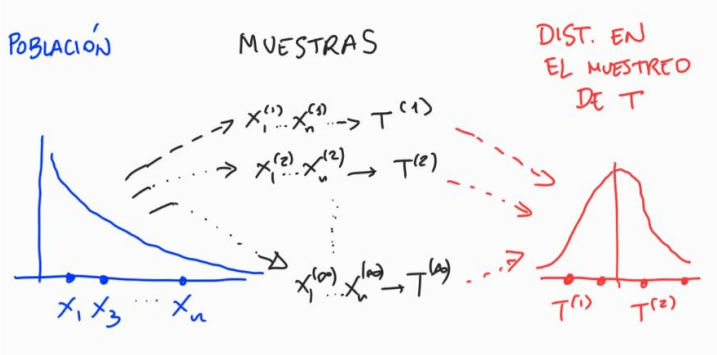
$$\mu_1 = \mu_0 \frac{1/\sigma_0^2}{1/\sigma_0^2 + n/\sigma^2} + \bar{x} \frac{n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}.$$

Métodos bayesianos y computación

- Las **familias conjugadas** para un modelo dado son familias paramétricas de distribuciones a priori tales que la distribución a posteriori pertenece a la misma familia paramétrica, cuando los datos siguen ese modelo.
- Si las distribuciones no son conjugadas, los problemas computacionales que presenta el cálculo de la distribución a posteriori y su esperanza pueden ser muy difíciles, especialmente si θ es un vector de alta dimensión.
- Se han desarrollado métodos numéricos basados en simulación de cadenas de Markov (**Gibbs sampling** y, más en general, **métodos MCMC (Markov chain Monte Carlo)**) que permiten extender la aplicación de los métodos bayesianos a modelos muy complejos con un número grande de parámetros.

Criterios para valorar un estimador

- Un estimador $\hat{\theta} = T(X_1, \ldots, X_n)$ es también una variable aleatoria cuya distribución se denomina **distribución en el muestreo** del estimador.
- Esta distribución informa de los valores que podemos esperar que tome $\hat{\theta}$ si dispusiéramos de muchas muestras de la misma población.
- La distribución en el muestreo determina la calidad de un estimador.



Sesgo y varianza

Una buena propiedad de un estimador es que no tenga tendencia sistemática a infraestimar o sobreestimar el parámetro.

Se dice que un estimador $\hat{\theta}$ es **insesgado** si $\mathrm{E}(\hat{\theta}) = \theta$, para todo $\theta \in \Theta$.

En el caso de que esto no ocurra el **sesgo** se define como $\mathrm{Sesgo}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$. Si el sesgo es positivo, hay una tendencia sistemática a sobreestimar el parámetro, y lo contrario si es negativo.

Otra buena propiedad de un estimador es no dar resultados muy diferentes para las distintas posibles muestras. Esto significa que es bueno que la **varianza** del estimador, $\mathrm{Var}(\hat{\theta})$, sea lo menor posible.

El **error cuadrático medio** tiene en cuenta tanto el sesgo como la varianza simultáneamente:

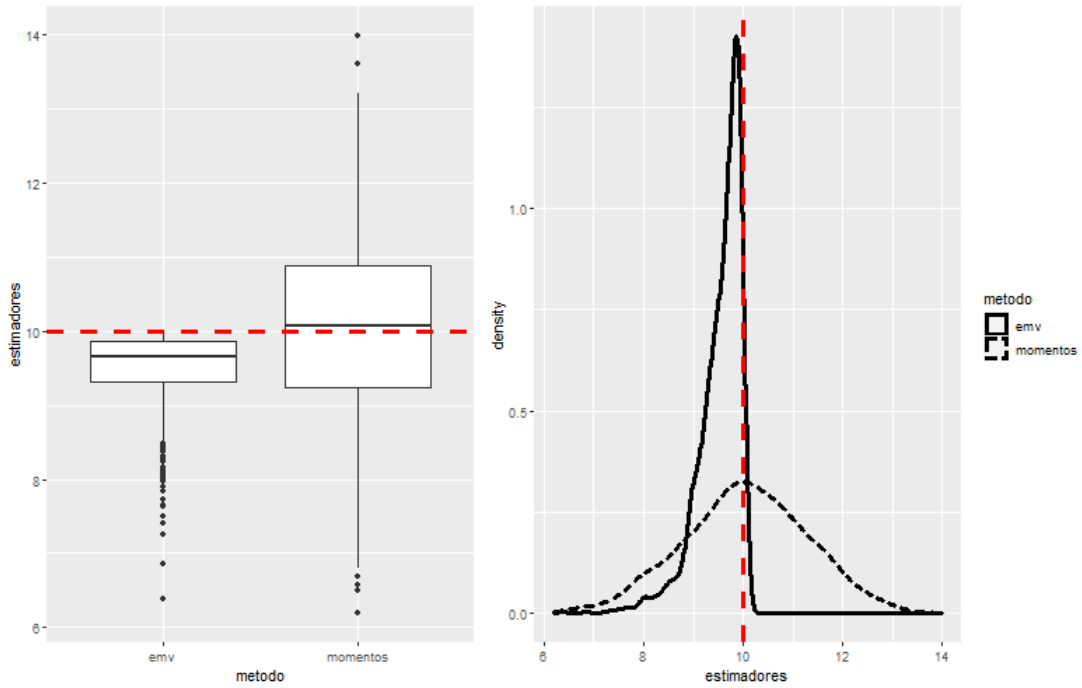
$$\mathrm{ECM}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2]$$

El ECM es igual al sesgo al cuadrado más la varianza.

Ejemplo: distribución uniforme

```
# Parámetros -------------------------------------------
theta <- 10  # valor verdadero del parámetro
n <- 20   # tamaño muestral
m <- 1000 # número de muestras

# Genera los datos -------------------------------------
set.seed(1234)   # para reproducir los resultados
muestras <- matrix(runif(n*m, 0, theta), n)

# Calcula estimadores ----------------------------------
emv <- apply(muestras, 2, max)
momentos <- 2*apply(muestras, 2, mean)
metodo <- gl(2, m, labels = c('emv', 'momentos'))
df <- data.frame(estimadores = c(emv, momentos), metodo =

# Gráficos ---------------------------------------------
cajas <- ggplot(df) +
  geom_boxplot(aes(x = metodo, y = estimadores)) +
  geom_hline(yintercept = theta, col = 'red', size = 1.1,

densidades <- ggplot(df) +
  geom_density(aes(x = estimadores, linetype = metodo), s
  geom_vline(xintercept = theta, col = 'red', size = 1.1,

cajas + densidades
```

Ejemplo: distribución uniforme



Otros criterios: consistencia, normalidad asintótica,...

- Los criterios asintóticos para evaluar un estimador se refieren a su comportamiento límite a medida que disponemos de más y más datos.
- Se dice que un estimador es **consistente** si su valor converge al del parámetro al aumentar el tamaño muestral. Dado que $\hat{\theta}$ es una sucesión de v.a. debemos considerar algún tipo de convergencia estocástica.
- Otra buena propiedad que puede tener un estimador es la **normalidad asintótica**, es decir, que la distribución límite del estimador sea aproximadamente normal para muestras grandes.
- Otros criterios: un estimador es **robusto** si no se ve muy afectado por la presencia de datos atípicos en la muestra.

Convergencia en probabilidad

Sea X_n una sucesión de variables aleatorias. Se dice que X_n **converge en probabilidad** a otra variable aleatoria X y se denota $X_n \to_p X$ si, para todo $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\{|X_n - X| > \epsilon\} = 0$$

Dado un estimador $\hat{\theta}$, se dice que es **consistente** para θ si $\hat{\theta} \to_p \theta$

Proposición. Si tanto el sesgo como la varianza de un estimador convergen a cero cuando $n \to \infty$, entonces el estimador es consistente.

Convergencia en distribución

Sea X_n una sucesión de variables aleatorias con funciones de distribución F_n. Se dice que X_n **converge en distribución** a otra variable aleatoria X con función de distribución F y se denota $X_n \to_d X$ si, para todo $x \in \text{Cont}(F)$, se verifica $\lim_{n\to\infty} F_n(x) = F(x)$, donde $\text{Cont}(F)$ es el conjunto de puntos en los que F es continua.

Relación entre las convergencias

$$X_n \to_P X \Rightarrow X_n \to_d X$$

Si la distribución límite es degenerada,

$$X_n \to_P \theta \Leftrightarrow X_n \to_d \theta$$

Teorema de la aplicación continua. Sea X_n una sucesión de v.a. tal que $X_n \to_d X$ y sea $g : \mathbb{R} \to \mathbb{R}$ una función continua. Entonces, $g(X_n) \to_d g(X)$.

Ley débil de los grandes números

Ley débil de los grandes números (LDGN). Sea X_n una sucesión de v.a.i.i.d. con media μ. Entonces,

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \to_p \mu.$$

La demostración cuando $\mathrm{Var}(X_i) = \sigma^2 < \infty$ se reduce a una aplicación elemental de la desigualdad de Chebychev:

$$\mathrm{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$$

Teorema central del límite

Teorema central del límite (TCL).} Sea X_n una sucesión de v.a.i.i.d. con media μ y varianza σ^2. Entonces,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to_d \mathrm{N}(0,1)$$

La conclusión del TCL es equivalente a

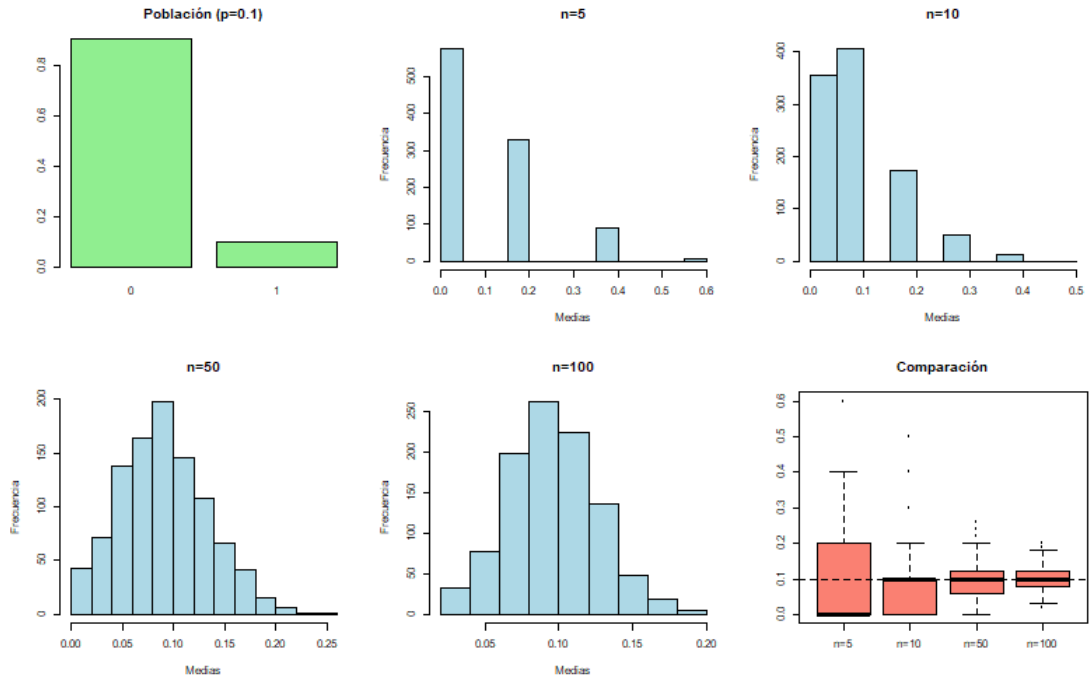
$$\sqrt{n}(\bar{X}_n - \mu) \to_d \mathrm{N}(0,\sigma^2)$$

El TCL da una aproximación de la distribución en el muestreo de la media:

$$\bar{X}_n \cong \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Promedios de 1000 muestras de tamaño n de una distribución de Bernoulli para distintos valores de n.

Ejemplo


Población (p=0.1)
n=5
n=10
n=50
n=100
Comparación
Frecuencia
Medias
n=5
n=10
n=50
n=100

Dos lemas útiles

La utilidad de estos dos resultados es extender las consecuencias de la LDGN y el TCL.

Lema de Slutsky. Sean X_n e Y_n dos sucesiones de v.a. tales que $X_n \to_d X$ e $Y_n \to_d \theta$, donde $\theta \in \mathbb{R}$. Sea $g : \mathbb{R}^2 \to \mathbb{R}$ una función continua. Entonces, $g(X_n, Y_n) \to_d g(X, \theta)$.

Método delta. Sean X_n una sucesión de v.a. tal que $n^b(X_n - \theta) \to_d X$ para $b > 0$ y $\theta \in \mathbb{R}$. Sea $g : \mathbb{R} \to \mathbb{R}$ una función derivable con derivada continua. Entonces,

$$n^b[g(X_n) - g(\theta)] \to_d g'(\theta)X$$

Algunas aplicaciones estadísticas

- Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con distribución uniforme en el intervalo $(0, \theta)$.
  - Determina el límite en distribución de $2\bar{X}_n$ y $\sqrt{n}(2\bar{X}_n - \theta)$.
  - Compara con el comportamiento asintótico de $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$.
- Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con distribución $\mathrm{B}(1, p)$. Determina el comportamiento asintótico de:
  - La proporción muestral: $\hat{p} = (X_1 + \cdots + X_n)/n$.
  - La proporción muestral estandarizada: $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$.
  - La proporción muestral estandarizada, pero usando en el denominador $\hat{p}$ en lugar de p: $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$.

Algunas aplicaciones estadísticas

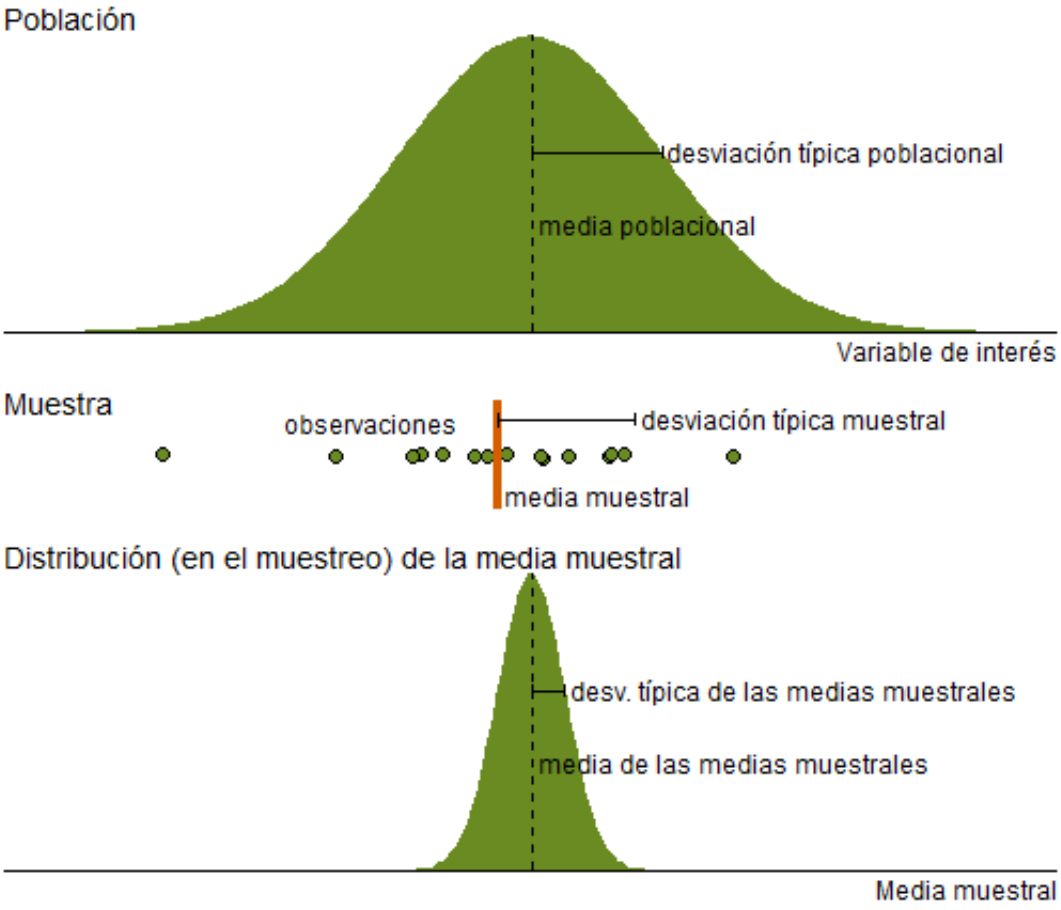
- Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con media μ, varianza σ^2 y $\mathrm{E}(X_i^4) < \infty$.
  - $S_n^2 \to_p \sigma^2$
  - $\sqrt{n}(S_n^2 - \sigma^2) \to_d \mathrm{N}(0, \sigma^4(\kappa - 1))$, donde $\kappa = \mathrm{E}[(X_i - \mu)^4]/\sigma^4$ es el llamado **coeficiente de curtosis**.
  - ¿Cuál es el comportamiento asintótico de $\sqrt{n}(S_n - \sigma)$?
- Supongamos que $X_1, X_2, \ldots$ son v.a.i.i.d. con distribución de Poisson de parámetro λ.
  - ¿Cuál es la distribución asintótica de $\sqrt{n}(\bar{X}_n - \lambda)$?
  - ¿Cuál es la distribución asintótica de $\sqrt{n}(2\sqrt{\bar{X}_n} - 2\sqrt{\lambda})$?

Propiedades de la media muestral

Dada una muestra de v.a.i.i.d. $X_1, \ldots, X_n$ de una población F, con $\mu = \mathrm{E}(X_i)$ y $\sigma^2 = \mathrm{Var}(X_i)$,

- $\bar{X}$ es insesgado para μ
- $\mathrm{Var}(\bar{X}) = \sigma^2/n$
- $\bar{X}$ es consistente para μ
- $\sqrt{n}(\bar{X} - \mu) \to_d \mathrm{N}(0, \sigma^2)$

Propiedades de la media muestral



Otras técnicas de inferencia: intervalos

- Sea una muestra $X_1, \ldots, X_n$ de una v.a. con función de distribución $F(\cdot;\theta)$, siendo $\theta \in \Theta \subset \mathbb{R}$ un parámetro desconocido.
- Sean dos estadísticos $a(X_1, \ldots, X_n)$ y $b(X_1, \ldots, X_n)$ con $a(X_1, \ldots, X_n) < b(X_1, \ldots, X_n)$ c.s. y un valor $\alpha \in (0, 1)$.
- Supongamos que se verifica

  $$\mathrm{P}_\theta\{a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)\} = 1 - \alpha, \text{ para todo } \theta.$$

- Para una realización concreta de la muestra, $x_1, \ldots, x_n$, se dice que $(a(x_1, \ldots, x_n), b(x_1, \ldots, x_n))$ es un **intervalo de confianza para θ con nivel de confianza $1 - \alpha$**.
- Lo denotaremos $\mathrm{IC}_{1-\alpha}(\theta)$.

Otras técnicas de infrencia: contrastes

- Una hipótesis (paramétrica) es una afirmación que se hace sobre uno o varios de los parámetros de la población.
- El objetivo de un contraste es decidir si los datos de una muestra aportan suficiente evidencia empírica para rechazar una hipótesis.
- La hipótesis que se pretende refutar se llama **hipótesis nula** (se denota H_0) y la contraria se llama **hipótesis alternativa** (se denota H_1).
- Construir un contraste es dar una regla que, dada una muestra, permita decidir si se rechaza o se acepta la hipótesis nula. El subconjunto de muestras para las que se rechaza la hipótesis nula se llama **región crítica**

Otras técnicas de infrencia: contrastes

- Si $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$, podemos expresar en general las hipótesis nula y alternativa como $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_1$
- Al aceptar o rechazar H_0 podemos cometer dos tipos de errores:
  - **Error de tipo I:** Rechazar H_0 cuando es cierta.
  - **Error de tipo II:** Aceptar H_0 cuando es falsa.
- La función de potencia del contraste definido por la región crítica R se define como $\beta(\theta) = \mathrm{P}_\theta(R)$.
- El **tamaño o nivel de significación de un contraste** es la máxima probabilidad de error de tipo I:

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

- El **p-valor** de un contraste para una muestra dada es **el ínfimo de los valores α para los cuales se rechaza la hipótesis nula a un nivel de significación α** con esa muestra.

Ejemplo

- Tenemos una muestra $X_1, \ldots, X_n$ de vaiid con distribución $\mathrm{N}(\theta, 1)$. Queremos contrastar $H_0 : \theta \leq 0$ frente a $H_1 : \theta > 0$. Consideramos la región crítica $R = \{\bar{X} > c\}$
- Calcula la función de potencia $\beta(\theta)$ como función de c y determina el valor que debe tener c para que el tamaño del contraste sea un valor prefijado $\alpha \in (0, 1)$.
- ¿Cuál es la mayor de las probabilidades de error de tipo II para el contraste anterior?